ARE REGISTERED AUTHORS MORE PRODUCTIVE?

Sarah Heeffer[1], Bart Thijs[2], and Wolfgang Glänzel[3]

[1] *sarah.heeffer@kuleuven.be*
KU Leuven, ECOOM and Dept. MSI, Leuven (Belgium)

[2] *bart.thijs@kuleuven.be*
KU Leuven, ECOOM and Dept. MSI, Leuven (Belgium)

[3] *wolfgang.glanzel@kuleuven.be*
KU Leuven, ECOOM and Dept. MSI, Leuven (Belgium)
Library of the Hungarian Academy of Sciences, Dept. Science Policy & Scientometrics, Budapest (Hungary)

Introduction

The identification of authors in bibliographic databases and their assignment to research universities, research institutions or companies is still one of the big challenges in Scientometrics at the micro and meso level. Correct author identification is indispensable, above all, in longitudinal studies on scientific careers, studies of researchers' mobility or in monitoring constitution and performance of research teams (Strotman & Zhao, 2012). Recently the large abstract and citation databases Web of Science (Thomson Reuters) and Scopus (Elsevier) have introduced their ResearcherID and Author ID, respectively. Both are supposed to uniquely identify scientific authors but experience has taught us that these IDs are not yet fully implemented and that errors and multiple assignments are not quite the exception to the rule.

The present study aims at a systematic analysis of the cleanness of ResearcherIDs, their acceptance by authors and their implementation in the mirror of national research output and subject-specific peculiarities as reflected by major science fields. Finally we have analysed in how far ResearcherIDs can be used to represent national and field-specific publication-activity patterns. The latter question is important to find reference standards for publication activity such as otherwise only known for citation indicators so far.

Data sources and data processing

In order to use a reasonable publication set we have selected seven countries from Europe and one country from Asia. These countries are Austria, Belgium, Germany, Hungary, Netherlands, China, Switzerland and UK. All 'citable' documents with at least one author from these countries and one or more authors with ResearcherID (RID) have been downloaded from the 2009–2011 volumes of the online version of Thomson Reuters' (TR) *Web of Science* (WoS). It should be stressed that the author with RID needs not necessarily be affiliated with an institution in the countries in question. After download these papers have been matched with all publications from these countries extracted from the WoS custom-data set

licensed at ECOOM. In a following step all RIDs have been uniquely assigned to countries on the basis of TR's affiliation tag. RID's from foreign countries have been removed from the national sets. All authors without RID have also been assigned to countries and – as far as possible – disambiguated on the basis of name and first initial and affiliation. After the cleaning process a certain amount of homonyms and synonyms still remains in the data set as well as some uncertainty about the authors' consequent and correct mention of their identifiers. All papers have been assigned to major fields on the basis of the Leuven-Budapest classification scheme. Papers can be assigned to more than one field or country due to journal assignment and co-authorship, respectively.

Table 1. Shares of RID authors and papers with RID authors per country [Data sourced from Thomson Reuters Web of Knowledge]

Country	Papers	A (%)	B (%)	C (%)
AUT	36272	45.7	12.1	27.1
BEL	53682	42.8	13.7	28.4
DEU	277524	41.2	15.4	22.9
HUN	17073	49.6	20.9	31.6
NLD	97625	45.1	19.2	30.2
CHN	423510	36.5	13.0	26.4
CHE	69958	47.8	16.5	19.6
GBR	298857	48.8	12.5	27.7

Legend: A = Mean share of RID per paper, B = share of papers with RID, C = share of authors with RID

Methods and results

Researcher names associated with RIDs were matched with author names as they appear on the paper. This allowed us to identify some problems. First, RIDs are not only used by authors. Some institutes and author groups mark their publications by an RID. Some RIDs claim several papers while the researcher name does not match any of the authors. A RID is not always unique. Some authors have created and are using different RIDs to claim the same papers with these different RIDs. The overwhelming share of RIDs, however, seems to be created by individuals and used in a correct manner.

Table 1 displays the mean shares of authors with RID (A) and the share of papers (B) respectively authors (C) with an RID. On an average, 40%–50% of authors on a paper have a RID registration. In China we have found the lowest share, while Hungary and the UK have the highest one around 50%. National shares of papers with RID authors is much lower; it ranges between 12% and 21%. Here Hungary and the Netherlands are at the high end and the UK has jointly with Austria the lowest share. Similarly, Hungary and the Netherlands have the highest shares of registered authors but unlike the previous static, Germany and Switzerland form the low end here. Roughly one quarter to one third of all authors from the country selection use a RID registration. These effects are not the result of foreign collaboration since co-authors from other countries have been removed from the statistics.

Table 2. Mean publication activity of RID authors vs. authors in RID papers and all authors per country [Data sourced from Thomson Reuters Web of Knowledge]

Country	A	B	C
AUT	3.89	3.35	7.73
BEL	4.52	3.16	7.02
DEU	4.95	3.84	7.56
HUN	4.00	2.99	4.76
NLD	4.00	3.02	7.77
CHN	23.34	9.26	8.33
CHE	4.32	4.60	6.81
GBR	4.09	3.13	5.39

Legend: A = Mean activity of all authors, B = Mean activity of authors in RID papers, C = Mean activity of RID authors

The comparison of publication activity reveals other aspects of national patterns of RID use. The mean activity is certainly distorted by insufficient name disambiguation. Although the national statistics for all authors reflect similar activity for most countries (ranging from 4 to 5), China's extreme average activity points to identification issues.

Table 3. Mean publication activity of all authors (A) vs. RID authors (C) per major field [Data sourced from Thomson Reuters Web of Knowledge]

Field	A	C	Field	A	C
A	2.17	3.05	M	3.11	4.40
B	2.40	3.01	N	2.51	3.84
C	3.12	4.76	O	1.74	2.25
E	2.14	2.37	P	4.96	4.85
G	3.58	4.10	R	1.97	2.28
H	1.89	1.90	S	1.67	2.15
I	2.98	3.53	Z	2.48	3.53

Legend: A: agriculture & environment; B: biosciences (general, cellular & subcellular biology; genetics); C: chemistry; E: engineering; G: geosciences & space sciences; H: mathematics, I: clinical and experimental medicine I (general & internal medicine); M: clinical and experimental medicine II (non-internal medicine specialties); N: neuroscience & behavior; O: social sciences II (economical & political issues), P: physics; R: biomedical research; S: social sciences I (general, regional & community issues), Z: biology (organismic & supraorganismic level)

The mean activity of all authors in the RID set is generally somewhat lower but still in line with the activity of all authors. Here the Chinese value is more realistic. As expected, the activity of authors using RID (cf. column C in Table 2) is distinctly higher than the activity of all authors (except for China). However, China has still the highest activity, followed by the Netherlands, Austria and Germany. Of course, these values can be influenced by national publication profiles, therefor we have a look at subject-specific peculiarities of activity patterns before we have a closer look at the distribution of papers over authors using or not using RID. Because of the bias in the Chinese data, we have removed China in the following. Table 3 shows the mean activity (all authors vs. RID) for 12 major fields in the sciences and two fields in the social sciences. Again, the mean publication activity of RID authors generally exceeds that of the reference standard based on all authors. Physics forms the only exception. Also subject-specific peculiarities can be observed: mathematics and the social sciences have the lowest standards, followed by biomedical research and engineering. The deviation of the values presented in Table 3 from those in Table 2 are caused by the 'multidisciplinarity' of authors: RID authors are active in 2.5 fields on an average, while all authors in about 2.2 fields.

The mean activity of all authors in all fields combined amounts to 4.71, that of RID authors 6.87. Similarly, the corresponding share of authors with one paper amounts to 43.1% and 21.7%, respectively. Furthermore, RID authors are more productive at the high end of the distribution. The distribution is plotted in Figure 1. It goes without saying that the two distributions are distinctly different and it needs no further significance test.

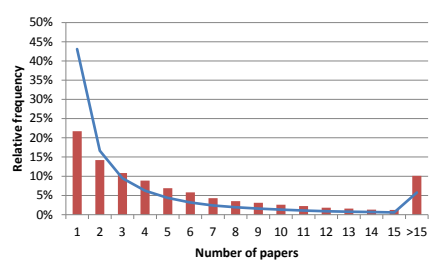

Figure 1. Relative frequency of publication activity of RID authors (bars) vs. all authors (line). [Data sourced from Thomson Reuters Web of Knowledge]

Conclusions

The validity of name disambiguation for some countries like China proved to be beyond tolerance. Nevertheless, the results leave no doubt. The extent of RID registration is still low and differs among countries. We also found that authors with RID are usually more productive than others. RID might therefore not (yet) be used to derive reference standards for publication activity.

Reference

Strotman, A. & Zhao, D. (2012), Author name disambiguation: What difference does it make in author-based citation analysis? *JASIST*, 63(9), 1820–1833.